Collaboratories: Sharing of a Common Context

Reagan W. Moore
San Diego Supercomputer Center
moore@sdsc.edu

Managing Context

- Content - the digital entities created by applications or sensors or desktop publishing
- Context - the properties that are used to describe relevance, manage, and analyze content
  - Data management context - state information resulting from operations on digital entities
  - Data discovery context - descriptive metadata used for browsing, queries, and analysis
  - Data preservation context - properties that facilitate management of technology evolution, while preserving information about the creation of the digital entity
- All three contexts are strongly inter-related
- Future data management will need to simultaneously support all three contexts

Communities

- Data grids (share data)
  - Manage distributed data (data generation sources)
  - Persistent naming convention
  - Manage state information resulting from operations on content
  - Manage consistency between context and content
- Digital libraries (publish data)
  - Manage discovery on context
  - Manage application of digital library services on context (curation, description, arrangement, storage, access)
- Persistent archives (preserve data)
  - Manage distributed data (replication for disaster recovery)
  - Persistent naming convention
  - Manage preservation context resulting from preservation processes (accession, description, arrangement, preservation, access)
  - Manage consistency between preservation context and content

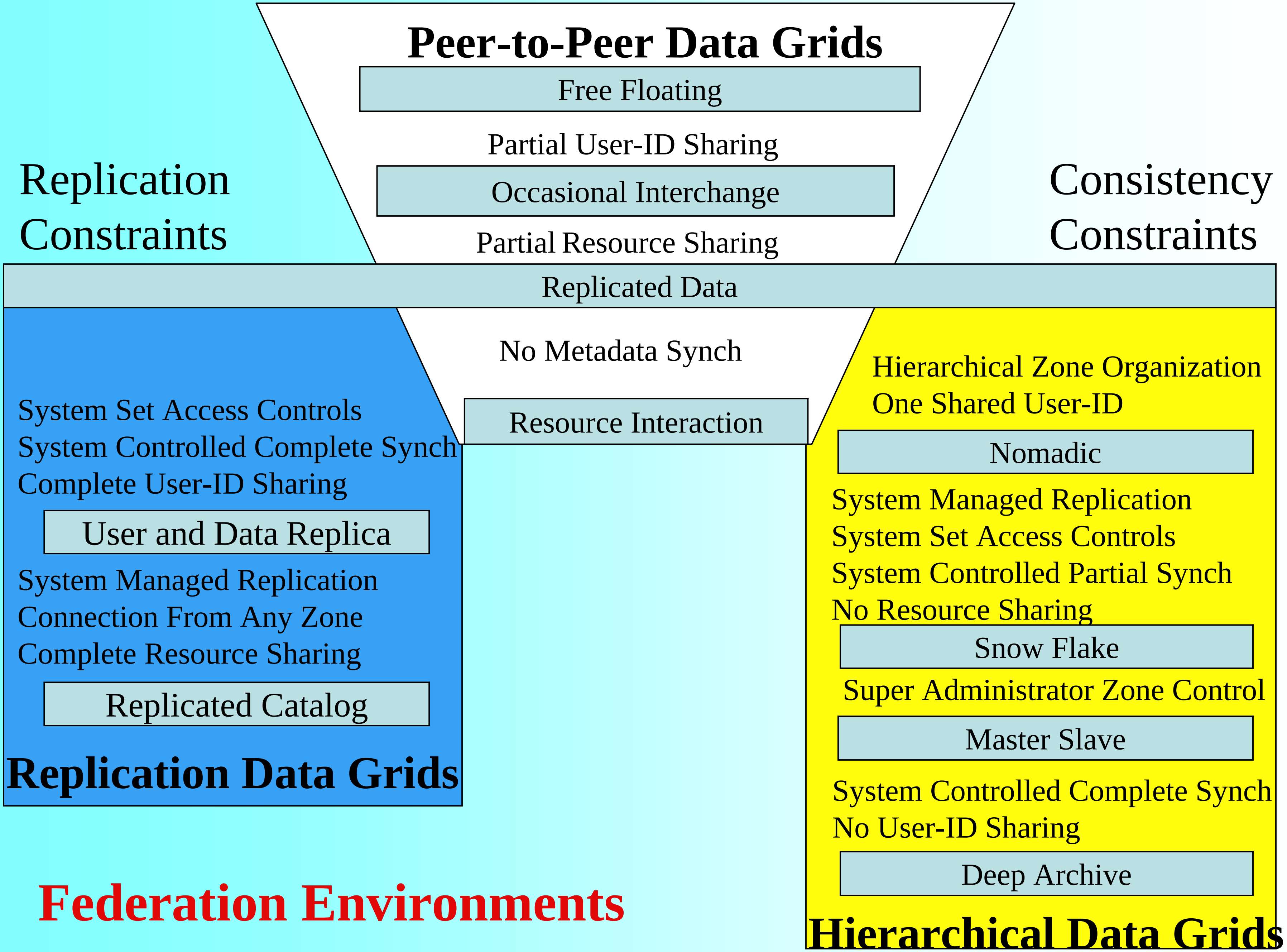
SRB Collections at SDSC

	As of 3/3/2004		As of 6/1/2004		As of 8/2/2004		
Project Instance	***Data_size (in GB)***	***Count (files)***	***Data_size (in GB)***	***Count (files)***	***Data_size (in GB)***	***Count (files)***	***Users***
Data Grid							
Digsky	45,939	8,685,572	51,380	8,690,003	51,380	8,690,003	80
NPACI	13,700	4,050,863	16,782	4,631,819	18,220	4,730,063	380
Hayden	7,835	60,001	7,201	113,600	7,201	113,600	178
SLAC -JCSG	3,432	446,613	4,161	551,918	4,898	617,374	47
LDAS/SALK	2,002	14,427	3,390	15,547	7,160	20,437	66
TeraGrid	22,563	452,868	58,228	481,489	94,203	704,493	3,077
BIRN	892	2,472,299	5,123	3,295,296	5,518	3,477,841	160
Digital Library							
DigEmbryo	720	45,365	720	45,365	720	45,365	23
HyperLter	215	5,110	224	5,166	241	7,065	35
Portal	1,610	46,278	1,690	46,011	1,767	48,513	384
AfCS	236	42,987	438	54,706	562	54,407	21
NSDL/SIO Exp	1,217	193,888	1,578	518,261	2,062	750,684	27
Transana	92	2,387	92	2,387	92	2,387	26
SCEC	12,311	1,730,432	14,738	1,735,900	25,715	1,753,458	56
UCSDLib	127	202,445	127	202,445	127	202,445	29
Persistent Archive							
NARA/Collection	72	82,192	63	81,191	63	81,191	58
NSDL/CI	1,529	12,658,072	2,445	18,491,862	3,203	23,559,785	122
TOTAL	114 TB	31 million	168 TB	39 million	223 TB	44 million	4757

** Does not cover data brokered by SRB spaces administered outside SDSC.
Does not cover databases; covers only files stored in file systems and archival storage systems
Does not cover shadow-linked directories

Common Context Management

- Common name spaces
  - Files / Users / Resources / Metadata
  - Collection based data management
- Infrastructure independence
  - Mapping from persistent name to location, descriptive metadata
  - Access controls on data and metadata, replication, aggregation
- Consistency
  - Audit trails, synchronization flags, write locks, metadata update
- Scalability of metadata management
  - Bulk operations, single application generates
    a million files, 100 TBs
- Federation of digital libraries
  - Integration of data management systems
    across sites, projects, disciplines, agencies



Peer-to-Peer Data Grids

Free Floating

Partial User-ID Sharing

Occasional Interchange

Partial Resource Sharing

Replication Constraints

Consistency Constraints

Replicated Data

No Metadata Synch

Resource Interaction

System Set Access Controls
System Controlled Complete Synch
Complete User-ID Sharing

User and Data Replica

System Managed Replication
Connection From Any Zone
Complete Resource Sharing

Replicated Catalog

Replication Data Grids

Hierarchical Zone Organization
One Shared User-ID

Nomadic

System Managed Replication
System Set Access Controls
System Controlled Partial Synch
No Resource Sharing

Snow Flake

Super Administrator Zone Control

Master Slave

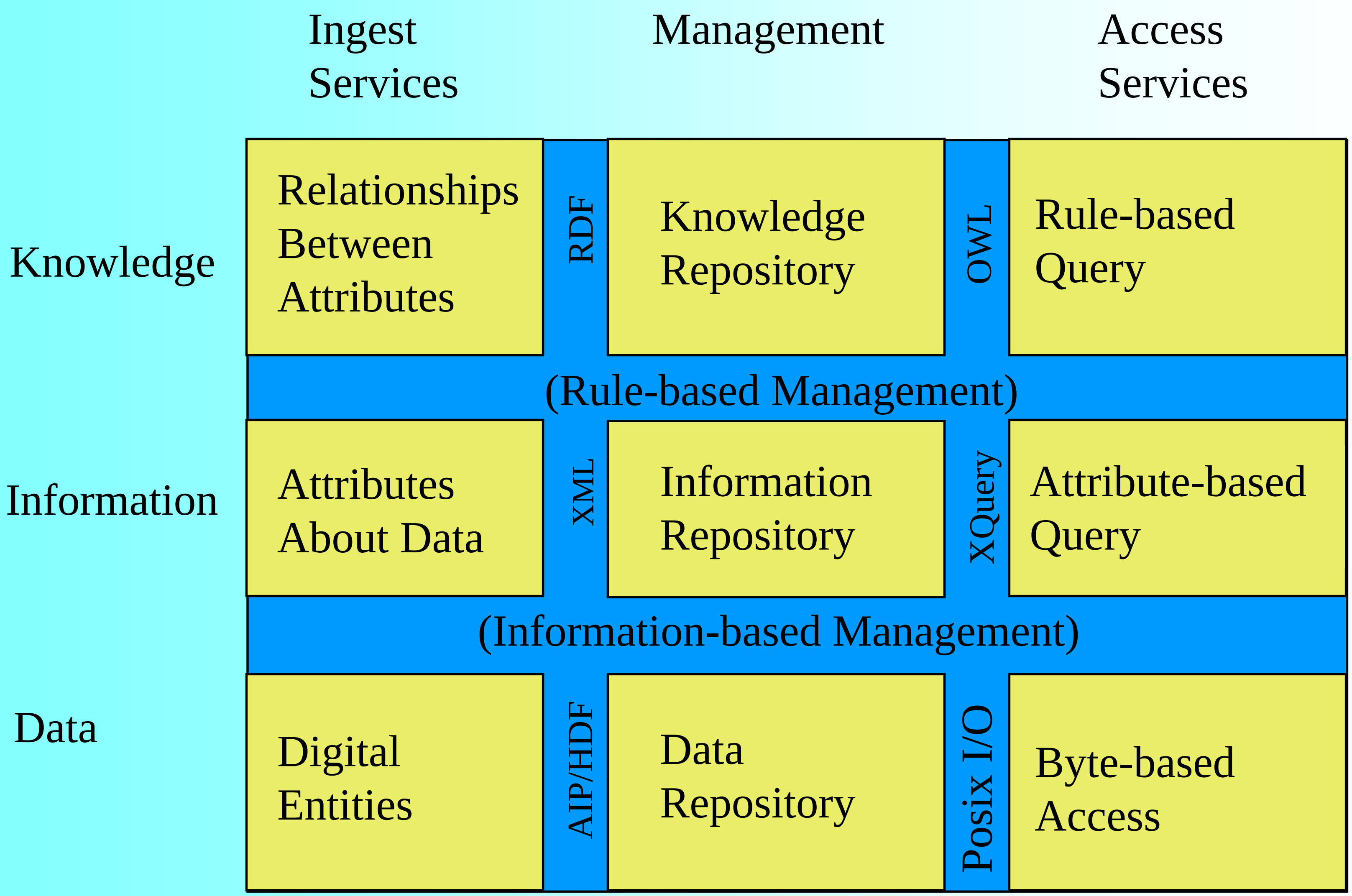
System Controlled Complete Synch
No User-ID Sharing

Deep Archive

Hierarchical Data Grids

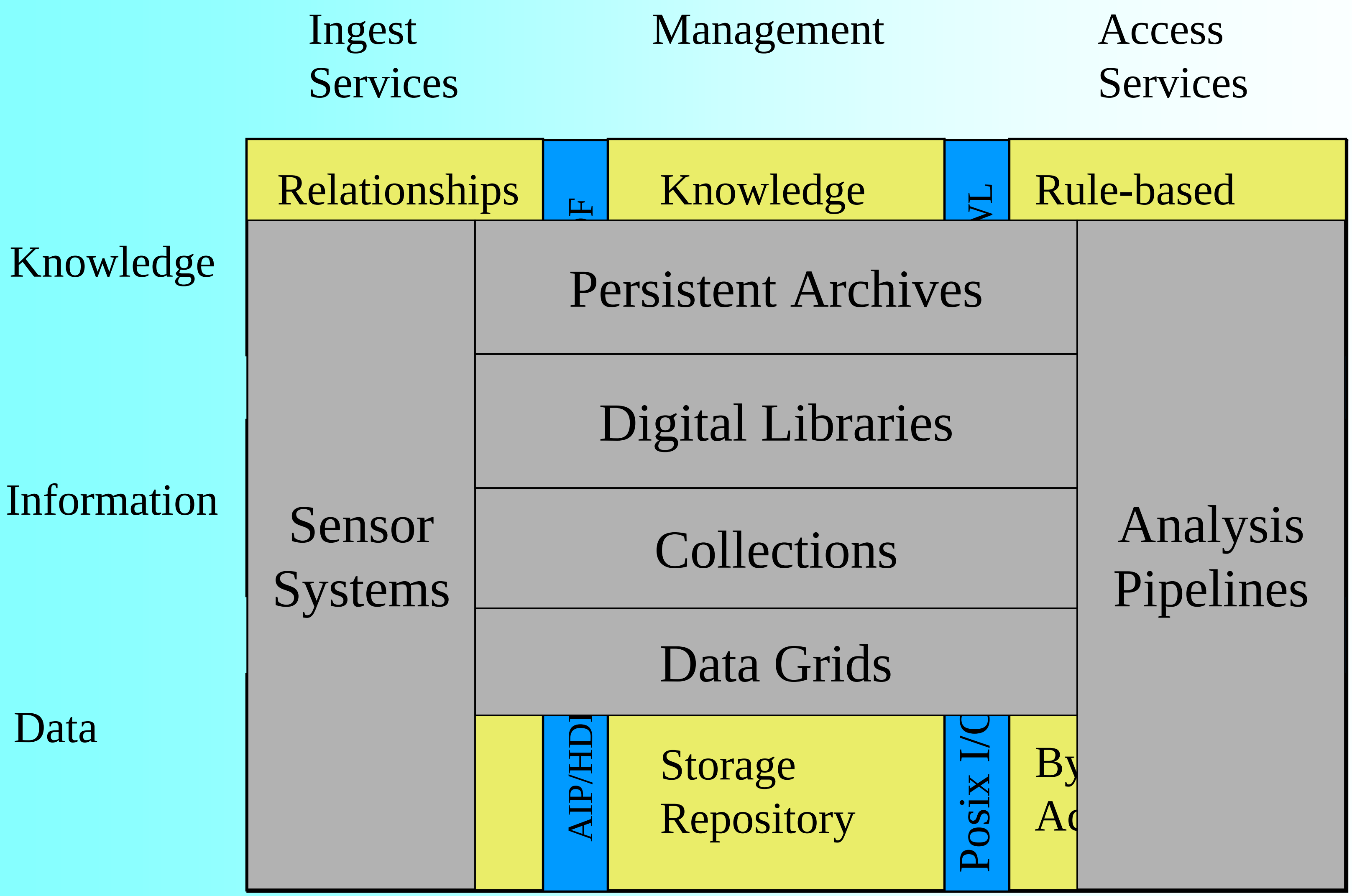
Federation Environments

Knowledge Based Data Management



	Ingest Services		Management		Access Services
Knowledge	Relationships Between Attributes	RDF	Knowledge Repository	OWL	Rule-based Query
	(Rule-based Management)				
Information	Attributes About Data	XML	Information Repository	XQuery	Attribute-based Query
	(Information-based Management)				
Data	Digital Entities	AIP/HDF	Data Repository	Posix I/O	Byte-based Access

Knowledge Based Data Management



Ingest Services

Management

Access Services

Relationships

F

Knowledge

VL

Rule-based

Knowledge

Information

Data

Sensor Systems

Persistent Archives

Digital Libraries

Collections

Data Grids

Analysis Pipelines

AIP/HDI

Storage Repository

Posix I/C

By Ac